Supplementary Materials for

Comprehensive serological profiling of human populations using a synthetic human virome

George J. Xu, Tomasz Kula, Qikai Xu, Mamie Z. Li, Suzanne D. Vernon, Thumbi Ndung'u, Kiat Ruxrungtham, Jorge Sanchez, Christian Brander, Raymond T. Chung, Kevin C. O'Connor, Bruce Walker, H. Benjamin Larman, Stephen J. Elledge
Correspondence to: selledge@genetics.med.harvard.edu

This PDF file includes:

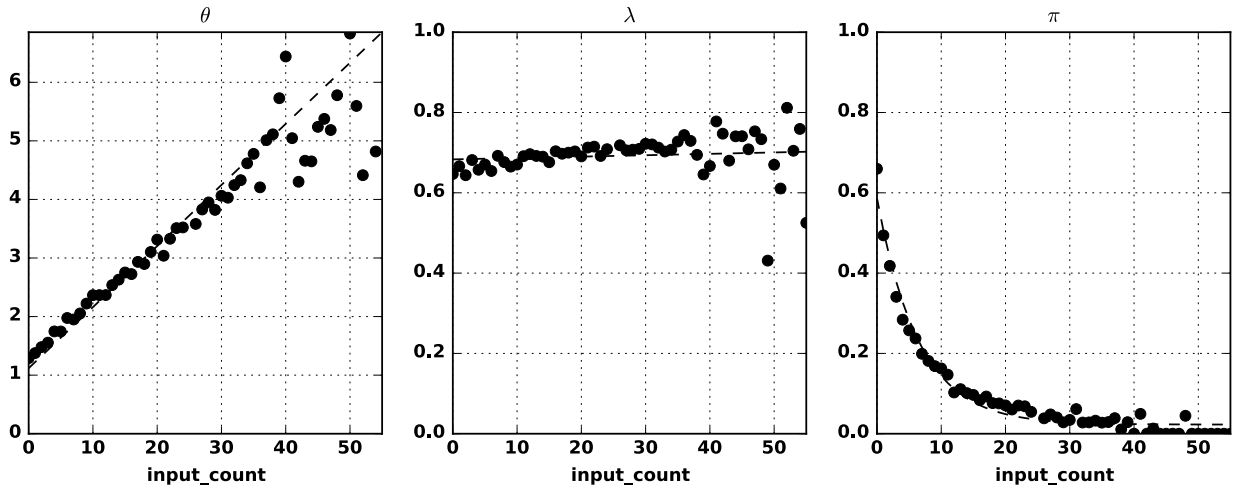

Fig. S1.

Zero inflated generalized poisson (ZIGP) parameters regressed on input count. Each scatter plot depicts the maximum likelihood estimates for the ZIGP parameters as a function of the input count (horizontal axis; see Materials and Methods). Dashed lines are least-squares linear regressions for θ and λ, and least-squares exponential regression for π.

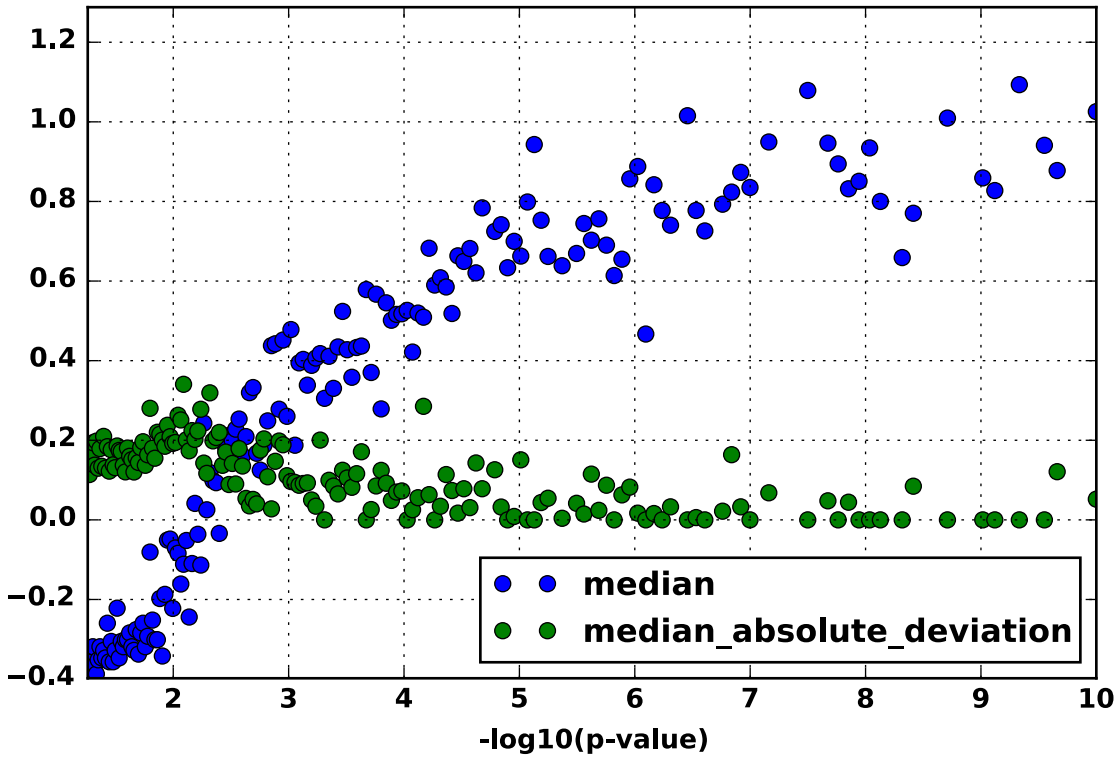

Fig. S2

Reproducibility threshold. Scatterplot for median and median absolute deviation of replicate 2 -log10(p-values) whose replicate 1 -log10(p-value) falls within the window whose left edge is shown on the horizontal axis (see Materials and Methods)..

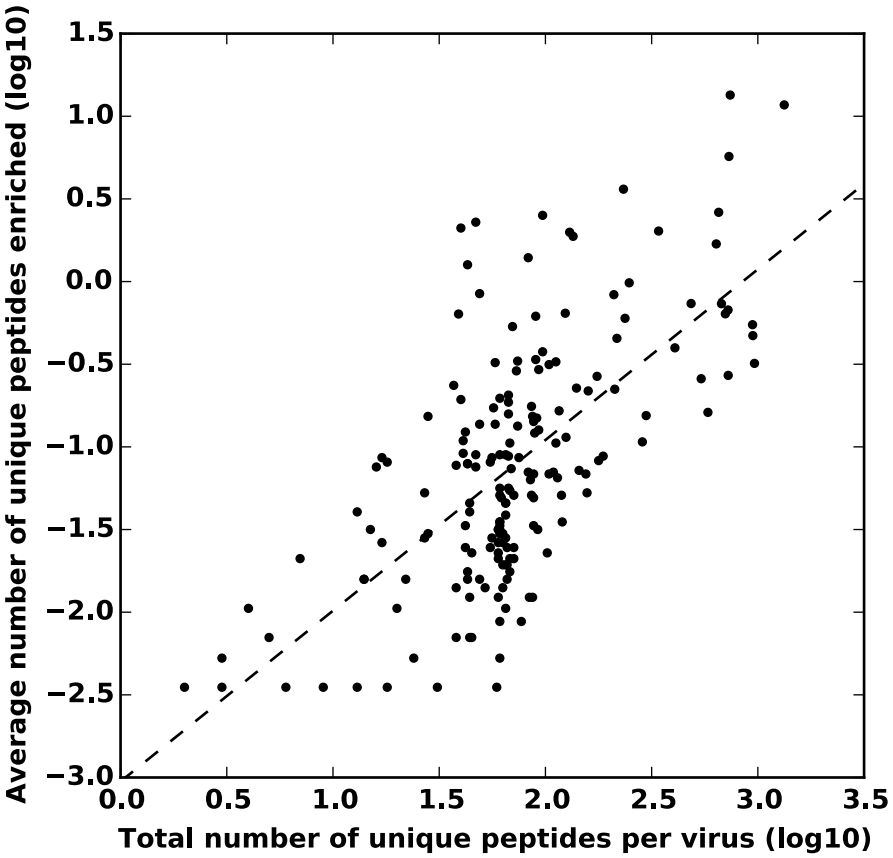

Fig. S3

Correlation between virus size and number of enriched peptides. Each dot on this log-log scatterplot is a virus. The horizontal axis corresponds to the size of the virus in number of peptides. The vertical axis corresponds to the average number of peptides enriched from the virus across all samples tested. The dashed line is a least-squares best-fit curve for the data.

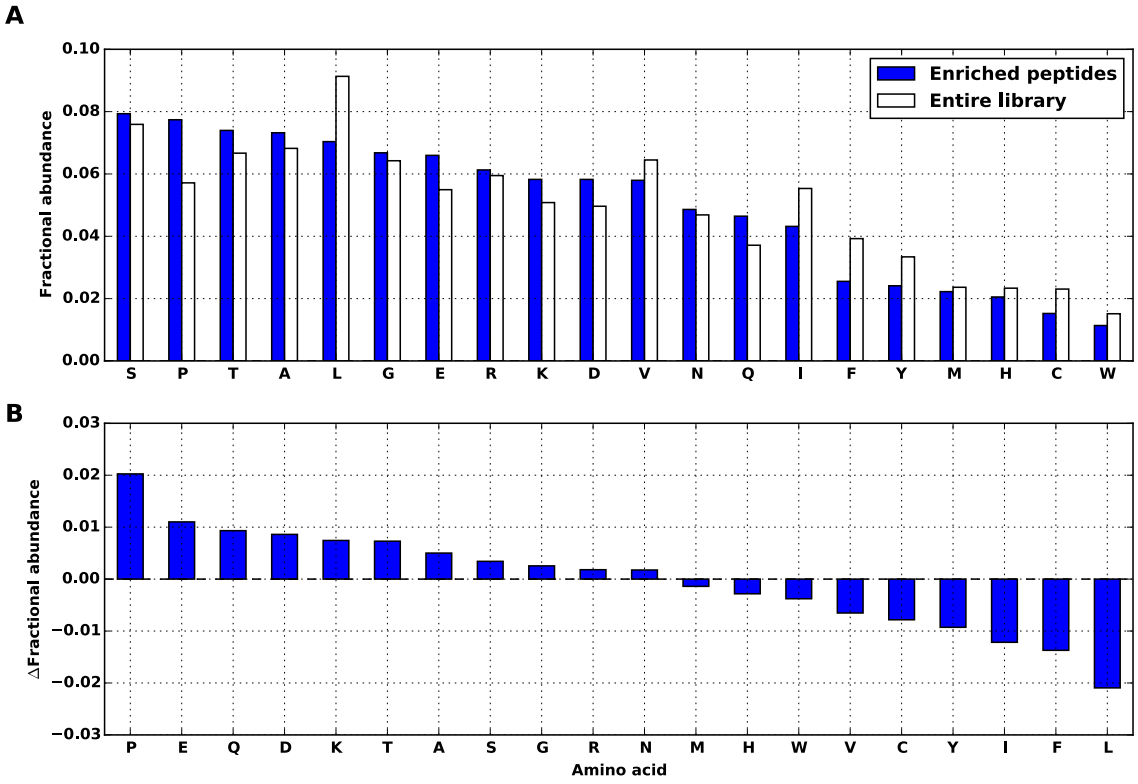

Fig. S4

Amino acid composition of enriched peptides. **(A)** Bar graph of the fractional abundance of each amino acid in the entire virome peptide library or peptides enriched in at least 2 samples. **(B)** Bar graph of the fractional abundance of each amino acid in peptides enriched in at least 2 samples subtracted by the abundance in the entire library.

Fig. S5-S11

Scanning mutagenesis identification of linear B cell epitopes in an immunogenic peptide from human viral proteins. Each row is a sample. Each column denotes the first mutated position for **(a)** single-, **(b)** double-, and **(c)** triple-alanine mutant peptides. The color intensity of each cell indicates the enrichment of the mutant peptide relative to the wild-type. For double-mutants, the last position is blank. The same is true for the last two positions for triple-mutants. Data shown are the mean of two replicates.

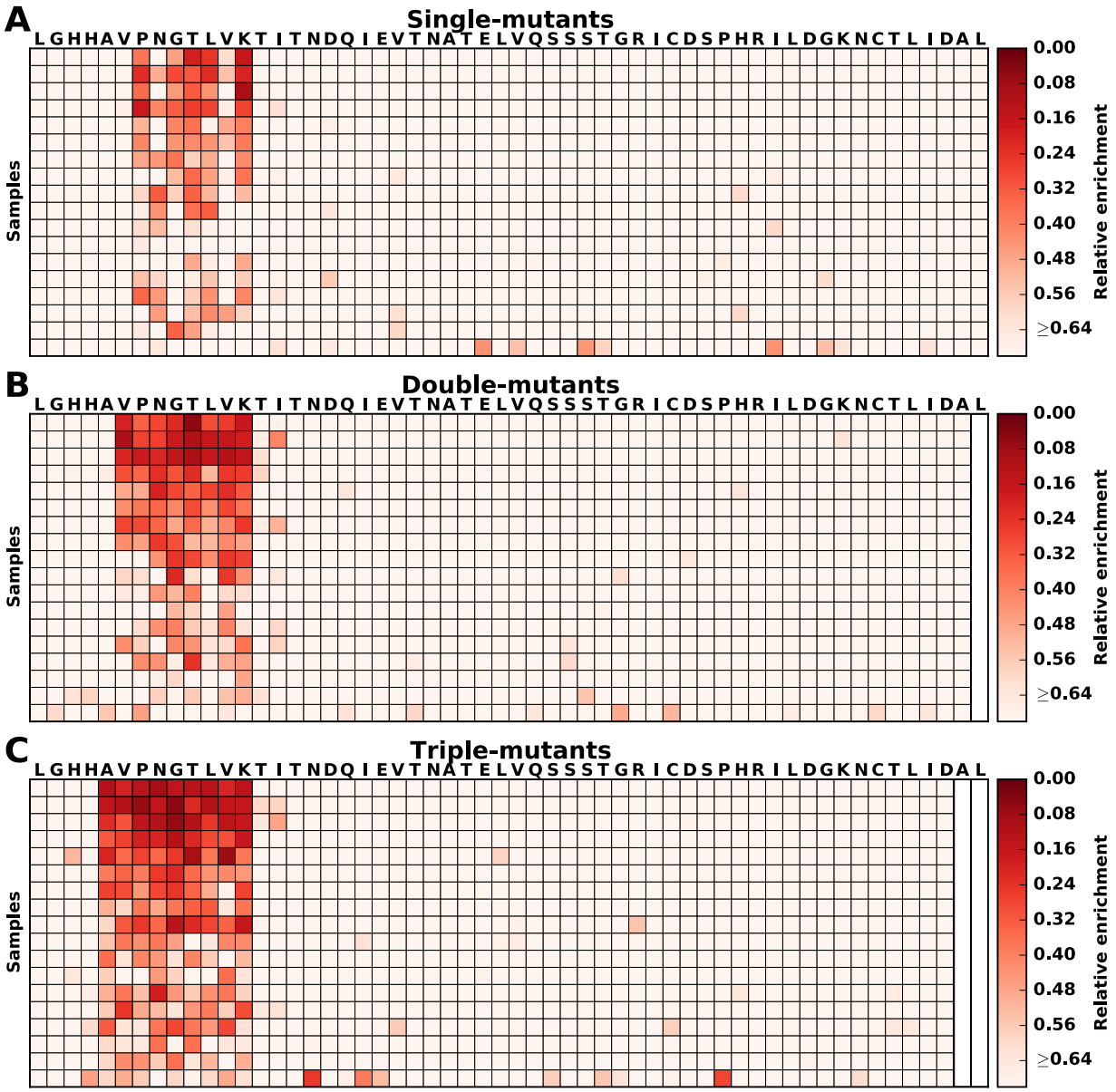

Fig. S5

Influenza A: hemagglutinin (UniProt ID: H8PET1, positions 1-56)

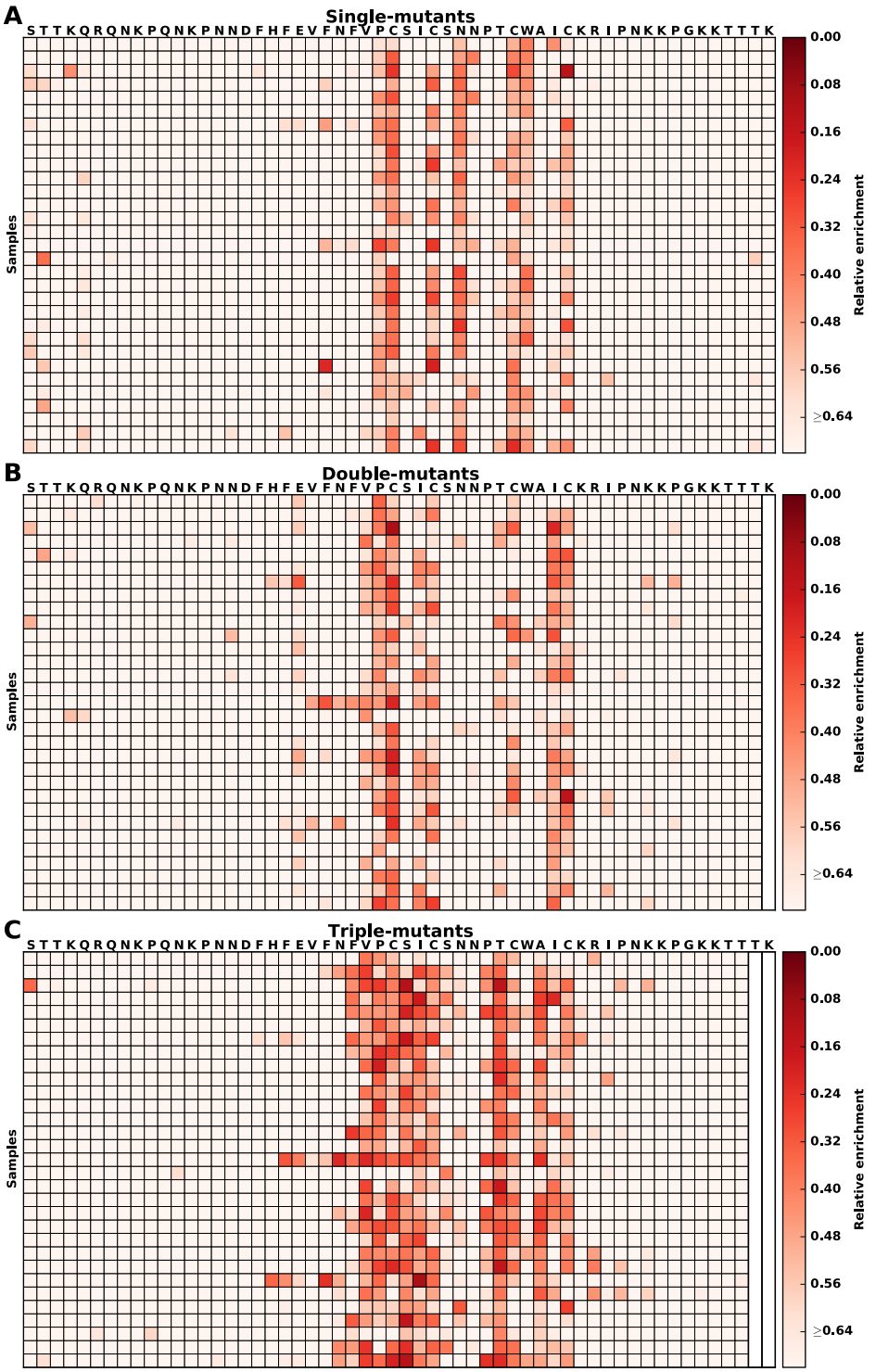

Fig. S6

Respiratory syncytial virus: attachment G glycoprotein (UniProt ID: P03276, positions 337-392)

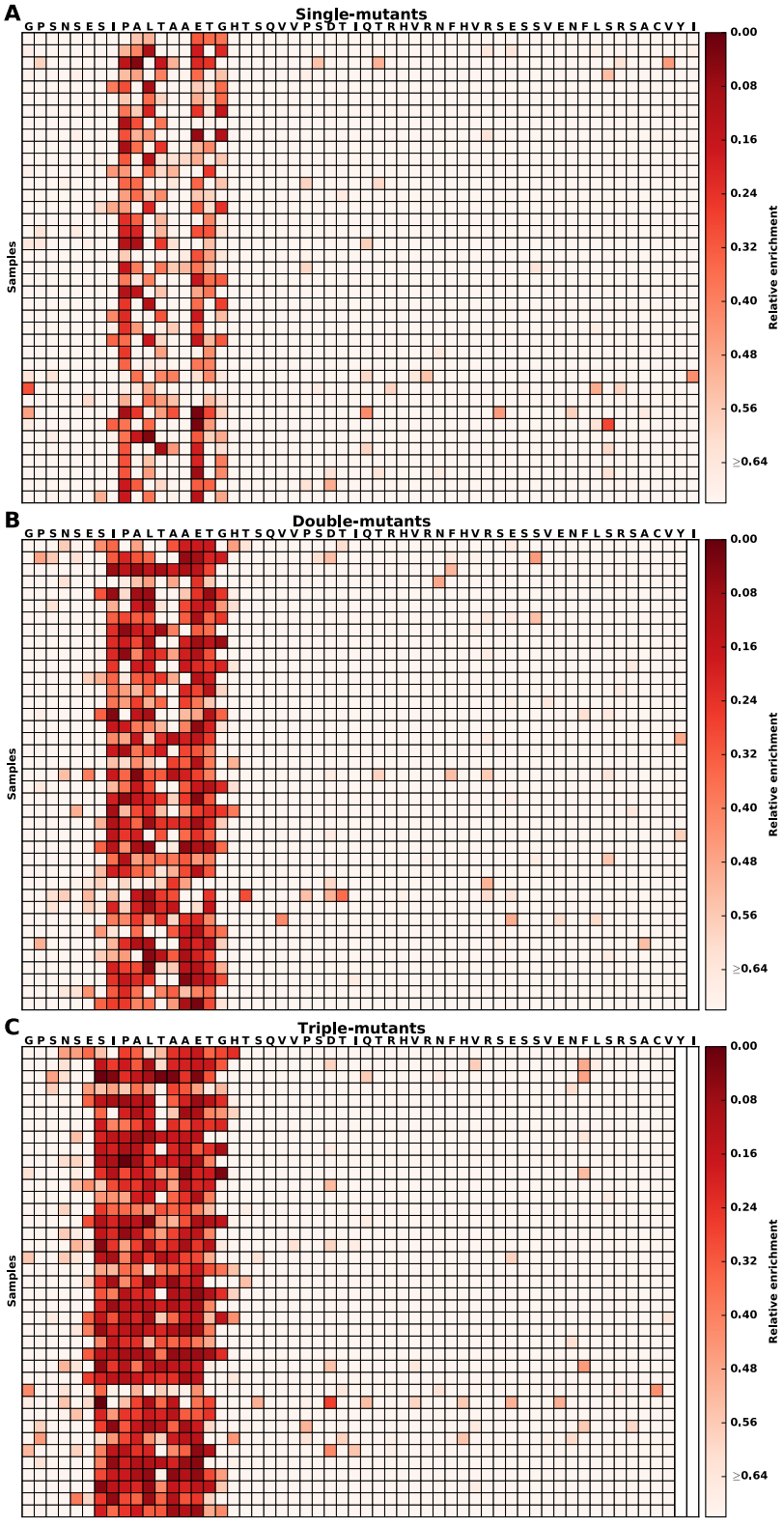

Fig. S7

Enterovirus B: genome polyprotein (UniProt ID: Q66474, positions 561-616)

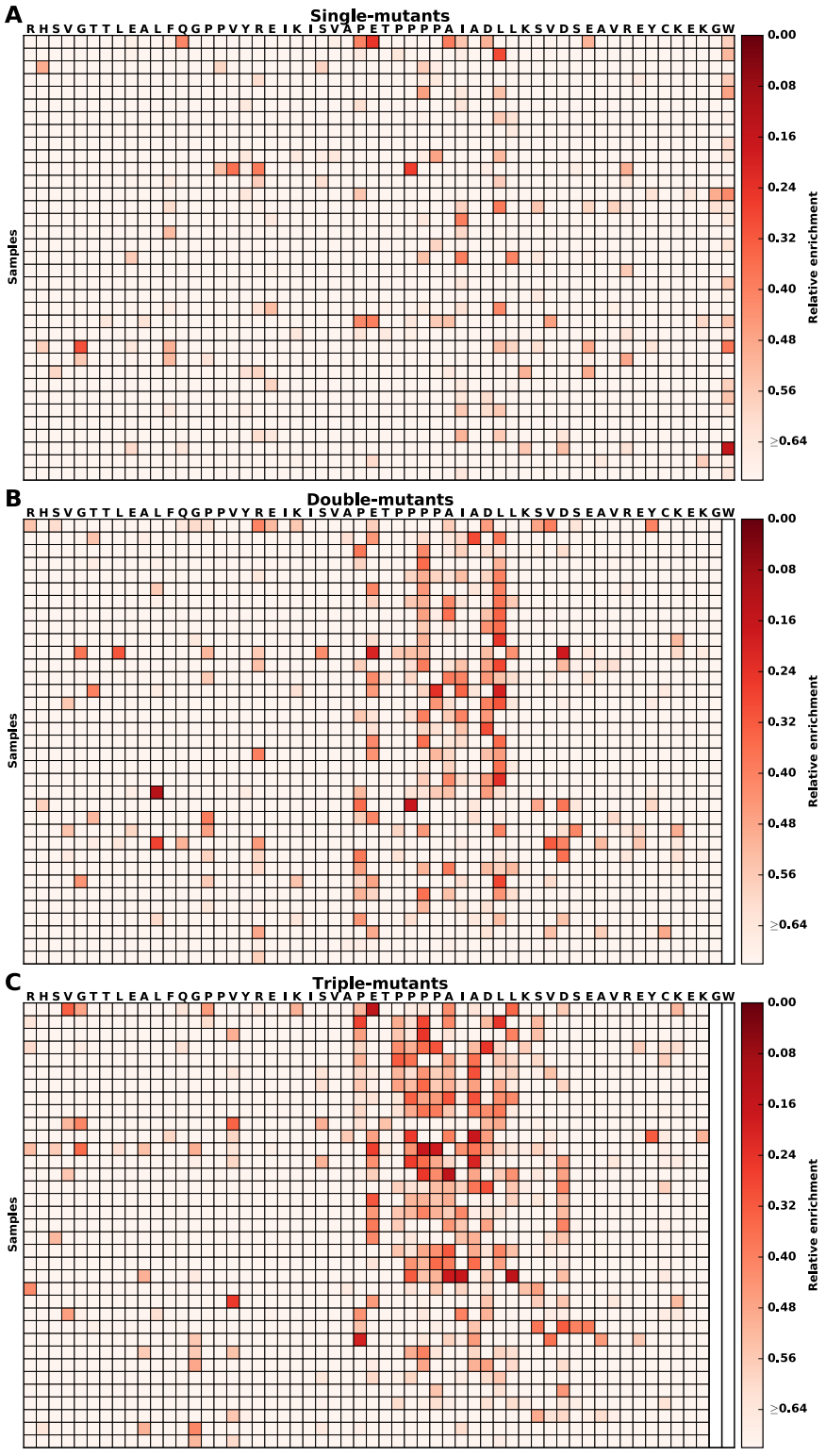

Fig. S8

Enterovirus B: genome polyprotein (UniProt ID: Q6W9F9, positions 1429-1484).

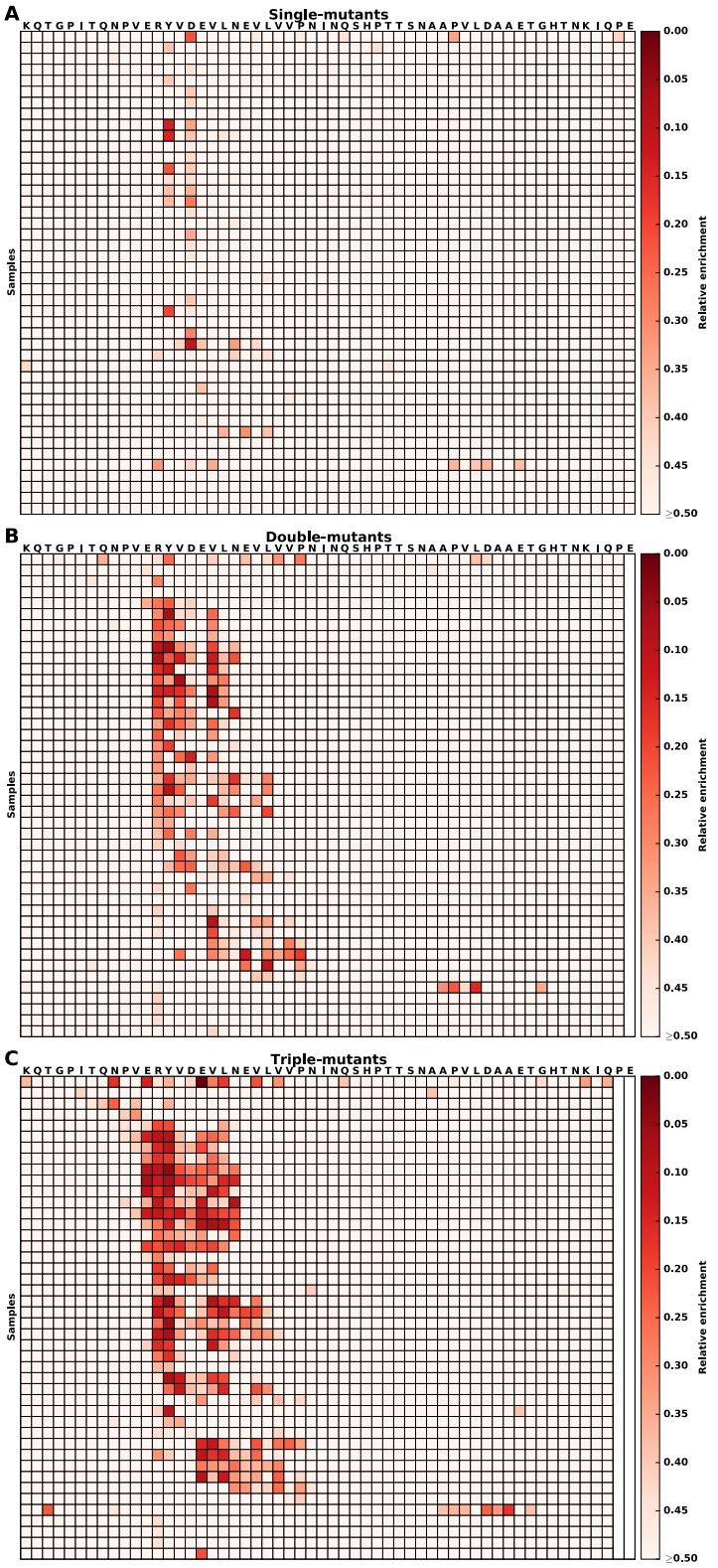

Fig. S9

Rhinovirus A: genome polyprotein (UniProt ID: Q82122, positions 561-616)

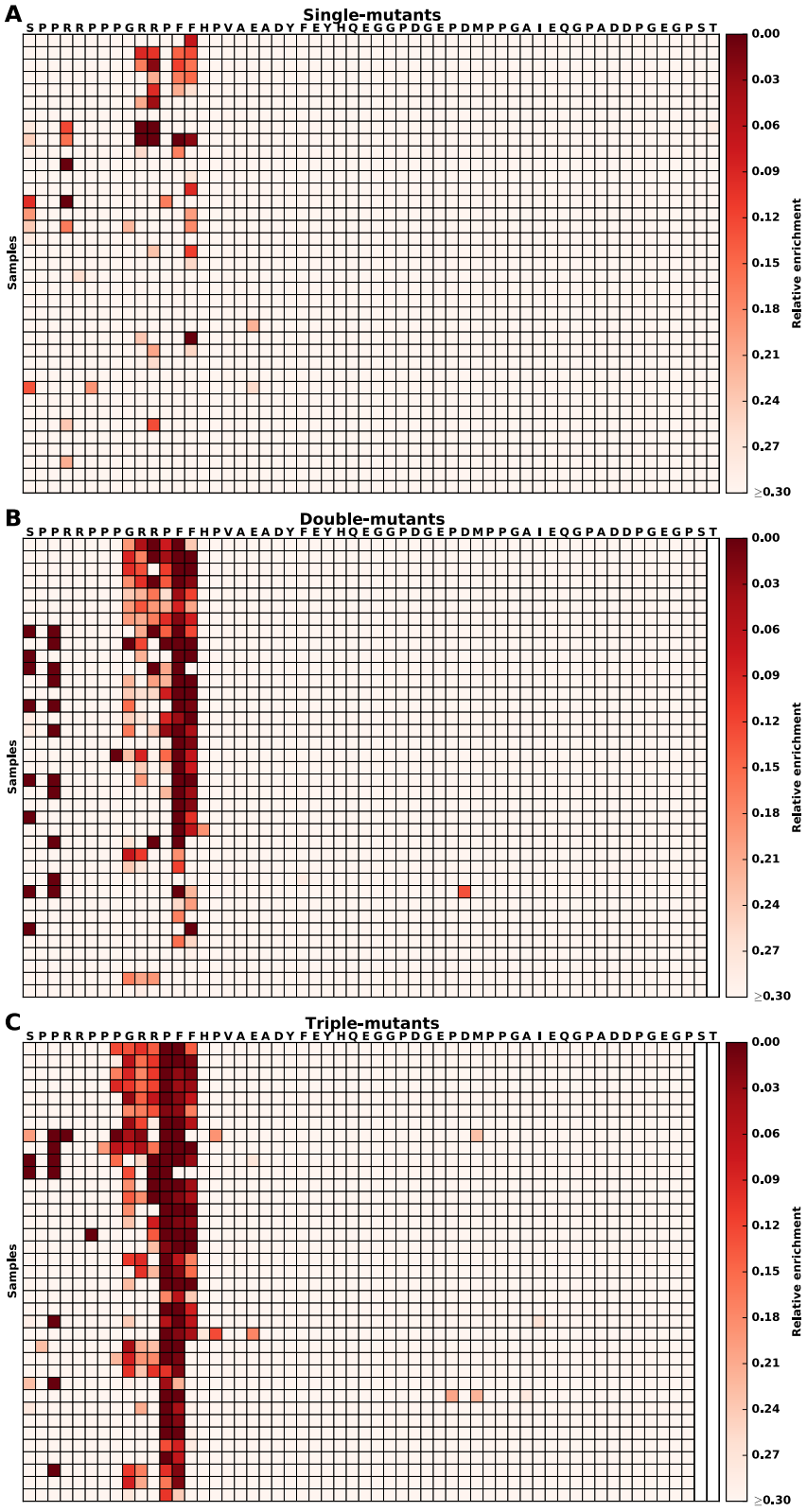

Fig. S10

Epstein-Barr virus: nuclear antigen 1(UniProt ID: Q1HVF7, positions 393-448).

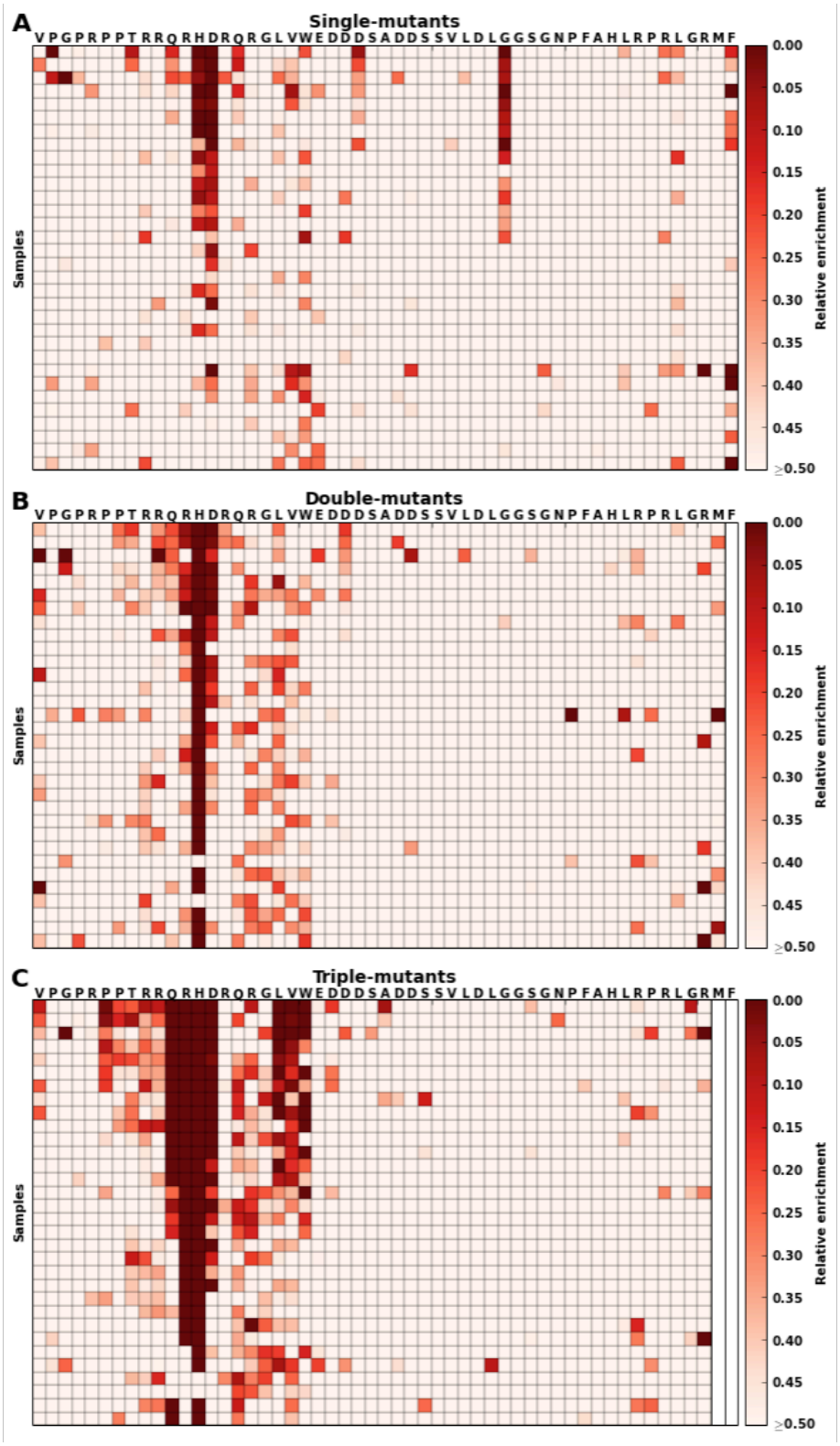

Fig. S11

Adenovirus C: precapsid vertex protein (UniProt ID: P03279, positions 533-585)

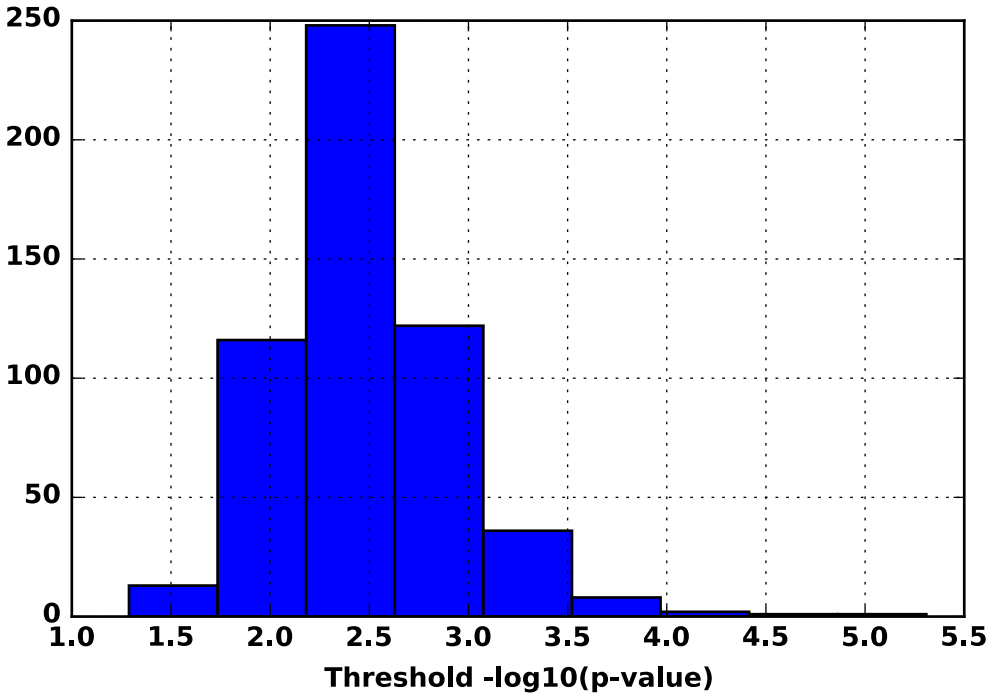

Fig. S12

Distribution of reproducibility threshold -log10(p-values). Histogram of the frequency of the reproducibility threshold -log10(p-values). The mean and median of the distribution are both approximately 2.3.

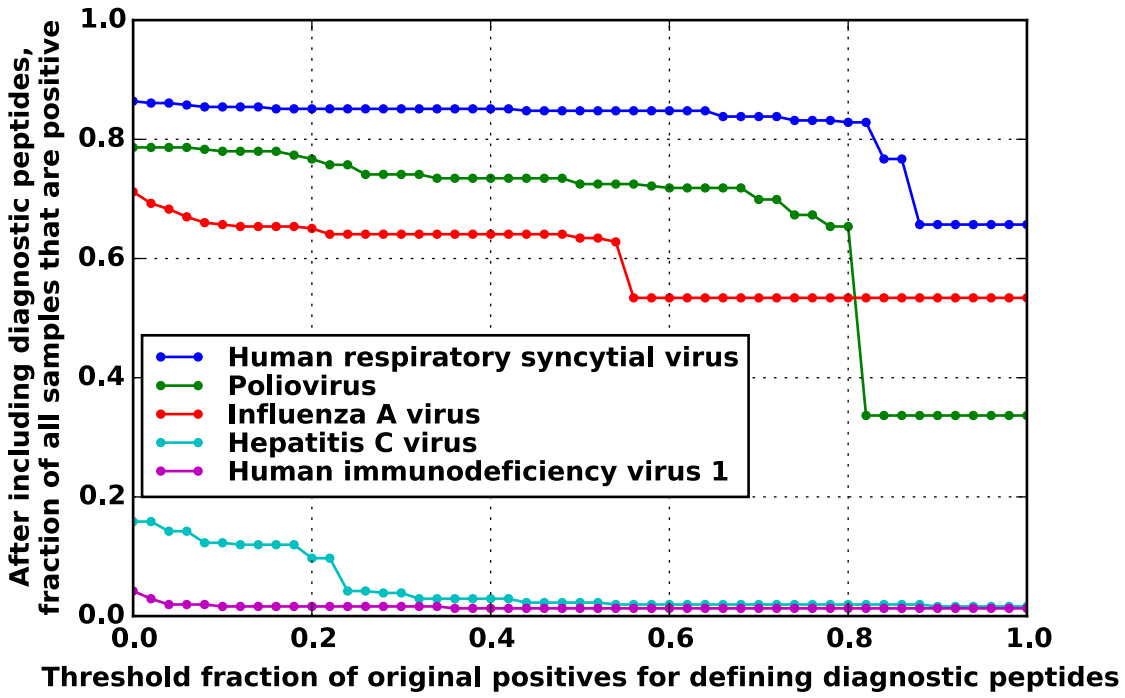

Fig. S13

Increased sensitivity after including samples targeting "diagnostic" peptides. For each virus, we examined all the samples that enriched multiple peptides that share a single epitope. If this epitope is "diagnostic" (i.e., recurrently targeted in at least a threshold fraction of the samples that were originally called positive for that virus), we considered the sample to be positive for that virus. The y-axis shows the fraction of samples that are considered positive after including these samples. The x-axis represents the minimum fraction of the original positive samples that must enrich a peptide for it to be considered diagnostic. Using a threshold of 30-70% significantly increases the rate of detecting respiratory syncytial virus without significantly increasing the rate of detecting hepatitis C and HIV, which should have low seroprevalence in this population (only samples from the United States that were not known HIV or HCV positives were included in this analysis).

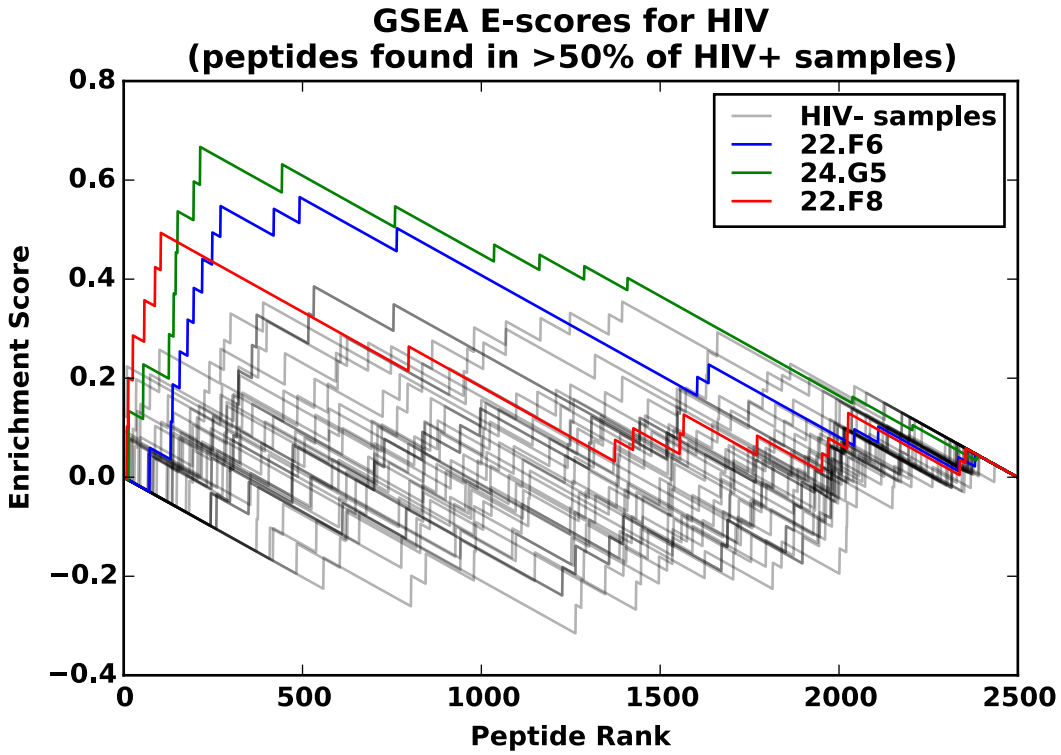

Fig. S14

Peptide set enrichment analysis for peptides containing recurrent epitopes in HIV samples. The analysis and graph are similar to the enrichment score calculation for the Gene Set Enrichment Analysis method. For each sample, the HIV peptides that did not pass our threshold for significantly enriched were ranked in descending order of $-\log_{10}$(p-value). A running sum was calculated by going down the list and, if the peptide was recurrently targeted in HIV (enriched in the majority of the HIV positive samples), the running sum was incremented by a value weighted by the $-\log_{10}$(p-value) of the peptide and normalized to 1 for all recurrent peptides. Otherwise, the running sum was decrement by a fixed value that was normalized to 1 for all non-recurrent peptides. The running sum is plotted for the 31 HIV negative samples (black lines) and for the HIV false negative samples (blue, green, and red lines). The maximum positive displacement of the running sum (enrichment score) is a measure of how significantly the set of peptides is enriched relative to the other HIV peptides.

Table S1.

Detection frequency of all viruses detected in at least 4 (>1%) of the 303 donors residing in the United States. Known HIV-positive and HCV-positive samples were excluded from this analysis. The "Detection Frequency" column shows the percentage of the 303 US samples that were positive for each virus. Of the samples that are positive for each virus, the "Above Minimum Threshold" column shows the percentage that enriched more unique peptides than just the minimum threshold for that virus (Fig S3), and the "Most Recurrent Peptide" column shows the percentage that enriched the most recurrent peptide for that virus. The "Number Unique Peptides Recurrent" column shows the number of unique peptides (peptides that do contain the identical subsequences of 7 amino acids or longer) from that virus that are enriched in at least 30% of the samples that are positive for that virus. The "Fraction Peptides Recurrent" column shows the total number of recurrent peptides from a virus divided by the number of all peptides from that virus.

	Detection Frequency	Above Minimum Threshold	Most Recurrent Peptide	Fraction Peptides Recurrent	Number Unique Peptides Recurrent
Human herpesvirus 4	87.1%	98.5%	87.4%	1.4%	13
Rhinovirus B	71.8%	52.7%	96.4%	5.0%	5
Human adenovirus C	71.8%	80.2%	71.6%	0.8%	4
Rhinovirus A	67.3%	59.1%	99.0%	4.6%	8
Human respiratory syncytial virus	65.7%	67.0%	86.2%	5.7%	4
Human herpesvirus 1	54.4%	87.5%	89.9%	1.1%	6
Influenza A virus	53.4%	57.0%	55.2%	0.1%	1
Human herpesvirus 6B	52.8%	66.3%	61.3%	0.7%	4
Human herpesvirus 5	48.5%	96.7%	95.3%	0.9%	19
Influenza B virus	40.5%	55.2%	51.2%	1.7%	4
Poliovirus	33.7%	40.4%	81.7%	2.0%	2
Human herpesvirus 3	24.3%	54.7%	77.3%	1.0%	4
Human adenovirus F	20.4%	17.5%	81.0%	0.4%	3
Human adenovirus B	16.8%	38.5%	71.2%	0.6%	3
Human herpesvirus 2	15.5%	75.0%	85.4%	0.7%	6
Enterovirus A	15.2%	12.8%	70.2%	2.3%	3
Enterovirus B	13.3%	7.3%	95.1%	3.3%	5
Mamastrovirus 1	9.4%	24.1%	55.2%	0.7%	2
Human herpesvirus 7	9.1%	42.9%	92.9%	0.4%	4
Norwalk virus	8.7%	25.9%	96.3%	1.2%	3
Human adenovirus D	8.4%	38.5%	50.0%	0.4%	3
Human parainfluenza virus 3	7.4%	21.7%	47.8%	1.6%	2
Cowpox virus	7.1%	9.1%	36.4%	0.1%	1
Human adenovirus A	6.5%	35.0%	55.0%	0.5%	2
Human metapneumovirus	5.2%	43.8%	43.8%	2.8%	4
Human coronavirus HKU1	4.5%	0.0%	42.9%	0.2%	3
Human herpesvirus 6A	4.2%	30.8%	46.2%	0.4%	4
Alphapapillomavirus 9	4.2%	30.8%	61.5%	0.5%	3
Human parvovirus B19	3.9%	25.0%	75.0%	1.5%	3
Aichivirus A	3.9%	33.3%	66.7%	2.6%	5
Hepatitis B virus	3.6%	9.1%	18.2%	0.0%	0
Betapapillomavirus 1	3.2%	0.0%	40.0%	0.1%	1
Influenza C virus	2.9%	33.3%	55.6%	0.2%	2
Human coronavirus NL63	2.9%	0.0%	55.6%	1.1%	3
Human herpesvirus 8	2.6%	50.0%	50.0%	0.5%	4
Rubella virus	2.6%	12.5%	50.0%	1.5%	2
Human adenovirus E	2.3%	14.3%	71.4%	0.5%	1
Hepatitis E virus	1.9%	0.0%	33.3%	0.4%	3
Torque teno virus	1.6%	0.0%	60.0%	0.9%	3
Hepatitis C virus	1.6%	80.0%	13.1%	0.0%	0
Measles virus	1.6%	20.0%	80.0%	2.2%	3
Alphapapillomavirus 10	1.6%	0.0%	80.0%	1.1%	3
Human parainfluenza virus 4	1.6%	0.0%	80.0%	6.3%	3
Eastern equine encephalitis virus	1.3%	0.0%	75.0%	0.7%	1
Rotavirus A	1.3%	0.0%	50.0%	0.1%	1

Table S2.

Modified algorithm applying more weight to diagnostic peptides shows improved detection of antibodies against respiratory syncytial virus (RSV). 60 patient sera were screened for RSV antibodies by ELISA and with VirScan. The concordance of the ELISA results with the initial and modified VirScan algorithms is shown in the tables.

Initial VirScan algorithm		RSV ELISA	
		Positive	Negative
RSV VirScan	Positive	37	1
	Negative	20	2

With diagnostic peptides		RSV ELISA	
		Positive	Negative
RSV VirScan	Positive	55	3
	Negative	2	0

Table S3.

Certain peptides are commonly targeted by the antibody response. We determined the peptide from each species of virus that was most frequently targeted in donors that were exposed to that virus. In each row, the frequency is the percentage of samples positive for the species of virus that had antibodies targeting the peptide sequence shown. The parent protein of the peptide is also listed.

Species	Protein	Peptide	%
Rhinovirus B	Genome polyprotein	QTVALTEGLGDELEEVIVEKTKQTVASISSGPKHTQKVPILTANETGATMPVLPSD	95%
Human herpesvirus 5	Envelope glycoprotein M	TASGEEVAVLSHHDSLESRRLREEEDDDDDEDFEDA	90%
Enterovirus B	Genome polyprotein	PFIQQEAKLQGEPGKAIESAISRVADTISSGPTNSEQVPALTAAETGHTSQVVPGD	86%
Human respiratory syncytial virus	Attachment glycoprotein	NKPSTKPRPKNPPKKPKDDYHFEVFNFVPCSICGNNQLCKSICKTIPSNKPKKKPT	85%
Human herpesvirus 4	Epstein-Barr nuclear antigen 1	SPPRRPPPGRRPFFHPVAEADYFEYHQEGGPDGEPDMPPGAIEQGPADDPGEGPST	81%
Human herpesvirus 1	Envelope glycoprotein D	RRHTQKAPKRIRLPHIREDDQPSSHQPLFY	80%
Norwalk virus	Genome polyprotein	LSSMAVTFKRALGGRAKQPPPRETPQRPPRPPTPELVKKIPPPPPNGEDELVVSYS	77%
Human adenovirus C	Pre-histone-like nucleoprotein	MTQGRRGNVYWVRDSVSGLRVPVRTRPPRN	74%
Enterovirus C	Genome polyprotein	QGALTLSLPKQQDSLPDTKASGPAHSKEVPALTAVETGATNPLAPSDTVQTRHVVQ	73%
Human herpesvirus 3	Envelope glycoprotein C	PDPAVAPTSAASRKPDPAVAPTSAASRKPDPAVAPTSAATRKPDPAVAPTSAASRK	72%
Human immunodeficiency virus 1	Envelope glycoprotein gp160	ERYLKDQQLLGIWGCSGKLICTTAVPWNASWSNKSLEQIWNNMTWMEWDREINNYT	60%
Influenza A virus	Hemagglutinin	LGHHAVPNGTLVKTITNDQIEVTNATELVQSSSTGRICDSPHRILDGKNCTLIDAL	42%

Supplementary Text

Estimating VirScan's specificity

Although we detected antibody responses to rare and highly virulent viruses such as Marburg and bat lyssavirus, they were found in less than 1% of the population (table S1), indicating that specificity is over 99% for these viruses, which is similar to the results in Table 1. Because we screened hundreds of sera for recognition of 206 virus species each, we performed the equivalent of approximately 100,000 individual tests, and eliminating such false positives altogether would require specificity of approximately 99.999% for each virus. Even with 99% specificity, a test will have 1% false positives, or approximately three per virus species for the 303 samples in population analyzed in Table S1.

In addition, 92 species of virus out of 206 were not detected in any samples from this population. Another 45 were detected in 3 or fewer samples. Assuming these are all false positives, which errs on the side of overestimating false positives, this analysis suggests that the specificity is 99.9%. While this is an imperfect estimate because we do not know how many of the detected positives are actually false positives, it gives an approximate estimate that argues the specificity is very high. No assay is perfect, and even highly optimized ELISAs for single viruses have some level of false positive, but our results give us a great deal of confidence in VirScan's specificity.

Differentially weighting recurrent peptides increases sensitivity

After discovering that certain epitopes are recurrently targeted, we examined whether we could apply this knowledge to improve the sensitivity of viral detection with VirScan. Recurrent epitopes make up a very small portion of a virus's proteome. On average, less than 1% of a given virus's proteome is targeted in more than 30% of samples positive for that virus. We hypothesized that samples showing a strong response to these recurrently targeted "diagnostic" peptides, which we defined as a peptide enriched in at least 30% of positive samples, are likely to be seropositive even if they do not meet our stringent cutoff requiring at least two non-overlapping enriched peptides. Thus, we introduced a modified criterion for calling a sample positive for a given virus that only requires one unique enriched peptide from the virus as long as the peptide is diagnostic (i.e., enriched in at least 30% of the samples that were originally called positive for that virus) and at least one other peptide that shares at least 7aa sequence homology was also enriched. The requirement for enrichment of two or more related peptides guards against potentially spurious technical enrichments.

We tested how this modified criterion affected our sensitivity and specificity in the known HCV positive and negative samples. In this set of samples, we had two false negatives, which had 11 and 14 enriched peptides, respectively, that were highly homologous and thus filtered down to one unique epitope. In both samples, this epitope corresponded to the N-terminus of the genome polyprotein, which is targeted in over 70% of the HCV positive samples. Thus, the modified criterion increases sensitivity for HCV to 100%. This modified criterion does not lead to increased false positives among known HCV negative samples nor does it significantly increase the rate of detecting HCV positive samples among the rest of the US samples (fig. S13).

We then tested how this modified criterion works on the known HIV positive and negative samples. Of the four false negative samples, one had enriched six related peptides targeted in 70-90% of the HIV positive samples and would be considered positive using the modified criterion. This relaxed criterion does not lead to increased false positives among known HIV negative samples nor does it increase the rate of detecting HIV positive samples among the rest of the US samples (fig. S13). The remaining three false negative samples did not significantly enrich a recurrently targeted peptide. However, upon further examination, we found that in two of the samples, although no single recurrent peptide was enriched, the set of recurrently targeted peptides were, as a group, enriched relative to other HIV peptides using a modified Gene Set Enrichment Analysis approach (fig. S14). These results suggest that these false negatives are due to low titers of anti-HIV antibodies that do not pass our stringent threshold for significance for any one peptide, but are significant when the set of homologous peptides are considered together. Once recurring peptides are identified for a given virus, this methodology could be used to develop a secondary analysis criterion for suspected false negatives, especially those that present with some but too few scoring peptides to meet the threshold for consideration as a positive.

We next turned our attention to respiratory syncytial virus (RSV), a virus for which our detected seroprevalence was lower than reported epidemiological rates, suggesting imperfect sensitivity of our assay. We tested 60 patient sera for antibodies to RSV by ELISA and found 95% were positive, above the reported sensitivity of the assay and consistent with near-universal exposure to this pathogen. Applying the modified criterion to these samples increased our rate of detection by VirScan from 63% to 97% (table S2). These data suggest that assigning more weight to recurrently targeted epitopes can enhance the sensitivity of VirScan and that the performance of the assay can be improved by screening known positives for a particular virus to discover these recurrently targeted epitopes.

Human Subjects Research Statement

The use of all samples for the purposes of this work was exempted by the Brigham and Women's Hospital Institutional Review Board (Protocol #: 2013P001337).